Chapter 12 Multivariate normal distributions

The multivariate normal is the most useful, and most studied, of the standard joint distributions in probability. A huge body of statistical theory depends on the properties of families of random variables whose joint distribution is at least approximately multivariate normal. The bivariate case (two variables) is the easiest to understand, because it requires a minimum of notation; vector notation and matrix algebra becomes necessities when many random variables are involved.

The general bivariate normal is often used to model pairs of dependent random variables, such as : the height and weight of an individual; or (as an approximation) the score a student gets on a final exam and the total score she gets on the problem sets; or the heights of father and son; and so on. Many fancy statistical procedures implicitly require bivariate (or multivariate, for more than two random variables) normality.

Bivariate normal

The most general bivariate normal can be built from a pair of independent random variables, X and Y, each distributed N(0, 1). For a constant ρ with $-1 < \rho < 1$, define random variables

U = X and $V = \rho X + \sqrt{1 - \rho^2} Y$

That is,

$$(U, V) = (X, Y)A$$
 where $A = \begin{pmatrix} 1 & \rho \\ 0 & \sqrt{1 - \rho^2} \end{pmatrix}$

Notice that $\mathbb{E}U = \mathbb{E}V = 0$,

$$var(V) = \rho^2 var(X) + (1 - \rho^2) var(Y) = 1 = var(U)$$

and

$$\operatorname{cov}(U, V) = \rho \operatorname{cov}(X, X) + \sqrt{1 - \rho^2} \operatorname{cov}(X, Y) = \rho$$

Consequently,

correlation(U, V) =
$$cov(U, V)/\sqrt{var(U)var(V)} = \rho$$

From Chapter 10, the joint density for (U, V) is

$$\frac{1}{|\det A|}f\left((u,v)A^{-1}\right),\,$$

where

$$f(x, y) = \frac{1}{2\pi} \exp\left(-\frac{x^2 + y^2}{2}\right)$$
 all x, y

The matrix A has determinant $\sqrt{1-\rho^2}$ and inverse

$$A^{-1} = \begin{pmatrix} \sqrt{1 - \rho^2} & -\rho \\ 0 & 1 \end{pmatrix} / \sqrt{1 - \rho^2}$$

Statistics 241: 16 November 1997

© David Pollard

Page 2

If $(x, y) = (u, v)A^{-1}$ then

$$x^{2} + y^{2} = (u, v)A^{-1}(A^{-1})'(u, v)'$$

= $(u, v) \begin{pmatrix} 1 & -\rho \\ -\rho & 0 \end{pmatrix} (u, v)'/(1 - \rho^{2})$
= $\frac{u^{2} - 2\rho uv + v^{2}}{1 - \rho^{2}}$

Thus U and V have joint density

$$\frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{u^2 - 2\rho uv + v^2}{2(1-\rho^2)}\right) \quad \text{for all } u, v.$$

•standard bivariate normal The joint distribution is sometimes called the STANDARD BIVARIATE NORMAL distribution with correlation ρ .

The symmetry of ψ in *u* and *v* implies that *V* has the same marginal distribution as *U*, that is, *V* is also N(0, 1) distributed. The calculation of the marginals densities involves the same integration for both variables.

When ρ equals zero, the joint density factorizes into

$$\frac{1}{\sqrt{2\pi}} \exp(-u^2/2) \frac{1}{\sqrt{2\pi}} \exp(-v^2/2)$$

which implies independence of U and V. That is, for random variables with a bivariate normal distribution, zero correlation is equivalent to independence. The equivalence for bivariate normals probably accounts for the widespread confusion between the properties of independence and zero correlation. In general, independence implies zero correlation, but not conversely.

<12.1> **Definition.** Random variables *S* and *T* are said to have a bivariate normal distribution, with parameters $\mathbb{E}S = \mu_S$, $\mathbb{E}T = \mu_T$, $\operatorname{var}(S) = \sigma_S^2$, $\operatorname{var}(T) = \sigma_T^2$, and correlation ρ if the standardized random variables $(S - \mu_S)/\sigma_S$ and $(T - \mu_T)/\sigma_T$ have a standard bivariate normal distribution with correlation ρ .

Problem 11.1 shows how to calculate explicitly the joint density for S and T.

Conditional distributions

The construction of U and V from the independent X and Y makes the calculation of the conditional distribution of V given U = u a triviality:

$$\rho X + \sqrt{1 - \rho^2} Y \mid X = x$$

has the distribution of $\rho x + \sqrt{1 - \rho^2} N(0, 1)$. That is,

$$V \mid U = u \sim N(\rho u, 1 - \rho^2)$$

<12.2>

The symmetry of the joint distribution of U and V implies that

$$U \mid V = v \sim N(\rho v, 1 - \rho^2),$$

a fact that you could check by explicit calculation of the ratio of joint to marginal densities:

$$\psi(u,v) \Big/ \int_{-\infty}^{\infty} \psi(u,v) \, du = \frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}} \exp\left(-\frac{(v-\rho u)^2}{2(1-\rho^2)}\right)$$

<12.3> **Example.** Let X denote the height (in inches) of a randomly chosen father, and let Y denote the height (in inches) of his son at maturity. Suppose each of X and Y has a $N(\mu, \sigma^2)$ distribution with $\mu = 69$ and $\sigma = 2$. Suppose also that X and Y have a bivariate normal distribution with correlation $\rho = .3$.

If Sam has a height of 74 inches, what would one predict about the ultimate height of his son Elmer?

Statistics 241: 16 November 1997

Page 3

In standardized units,

 $U = (X - \mu)/\sigma$ = Sam's standardized height, which happens to equal 2.5 $V = (Y - \mu)/\sigma$ = Elmer's standardized ultimate height.

By assumption, before the value of U was known, the pair (U, V) has a standard bivariate normal distribution with correlation ρ . From the analog of formula <12.2>,

$$V \mid U = 2.5 \sim N(2.5\rho, 1 - \rho^2)$$

In the original units,

Elmer's height | Sam's height = 74 inches $\sim N(\mu + 2.5\rho\sigma, (1 - \rho^2)\sigma^2) = N(70.5, 3.64)$

Notice that Elmer's expected height (given that Sam is 74 inches) is less than his father's height. This fact is an example of a general phenomenon called "regression towards the mean". The term REGRESSION, as a synonym for conditional expectation, has become commonplace in Statistics.

Multivariate densities

Random variables $X_1, X_2, ...$ are said to have a jointly continuous distribution with joint density function $f(x_1, x_2, ..., x_n)$ if

$$\mathbb{P}\{(X_1, X_2, \dots, X_n) \in A\} = \iint \dots \int \{(x_1, x_2, \dots, x_n) \in A\} f(x_1, x_2, \dots, x_n) \, dx_1 \, dx_2 \dots \, dx_n$$

for each subset A of \mathbb{R}^n . The density f must be nonnegative and integrate to 1 over \mathbb{R}^n .

It is convenient to write **X** for the RANDOM VECTOR (X_1, \ldots, X_n) , and **x** for the generic point (x_1, \ldots, x_n) in \mathbb{R}^n . Then the defining property for the joint density becomes

$$\mathbb{P}\{\mathbf{X} \in A\} = \int \{\mathbf{x} \in A\} f(\mathbf{x}) \, d\mathbf{x} \qquad \text{for } A \subseteq \mathbb{R}^n$$

where $\int \dots d\mathbf{x}$ should be understood as an *n*-fold integral.

<12.4> **Example.** If the random variables X_1, \ldots, X_n are independent, the joint density function is equal to the product of the marginal densities for each X_i , and conversely. The proof is similar to the proof for the bivariate case.

For example, if the $\{X_i\}$ are independent and each X_i has a N(0, 1) distribution, the joint density is

$$f(x_1, \dots, x_n) = \frac{1}{(2\pi)^{n/2}} \exp\left(-\sum_{i \le n} x_i^2/2\right) \quad \text{for all } x_1, \dots, x_n$$
$$= \frac{1}{(2\pi)^{n/2}} \exp(-\|\mathbf{x}\|^2/2) \quad \text{for all } \mathbf{x}$$

The distribution is denoted by $N(\mathbf{0}, I_n)$. It is sometimes called the "spherical normal distribution", because of the spherical symmetry of the density.

The methods for finding joint densities for random variables defined as functions of other random variables with jointly continuous distributions—as explained over the last two Chapters—extend to multivariate distributions. There is a problem with the drawing of *n*-dimensional pictures, to keep track of the transformations, and one must remember to say "*n*-dimensional volume" instead of area, but otherwise calculations are not much more complicated than in two dimensions.

Rotation of coordinate axes

The spherical symmetry of the density $f(\cdot)$ is responsible for an important property of multivariate normals. Let $\mathbf{q}_1, \ldots, \mathbf{q}_n$ be a new orthonormal basis for \mathbb{R}^n , and let

$$\mathbf{Z} = W_1 \mathbf{q}_1 + \ldots + W_n \mathbf{q}_n$$

© David Pollard

erandom vector

regression

Page 4

be the representation for \mathbf{Z} in the new basis.

<12.5> **Theorem.** The W_1, \ldots, W_n are also independent N(0, 1) distributed random variables.

In two dimensions, the assertion follows from the transformation formulae of Chapter 10. If the axes are rorated through an angle θ , then

$$W_1 = Z_1 \cos(\theta) + Z_2 \sin(\theta)$$
$$W_2 = -Z_1 \sin(\theta) + Z_2 \cos(\theta)$$

That is,

$$(W_1, W_2) = (Z_1, Z_2)A_\theta$$
 where $A_\theta = \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix}$

The matrix A_{θ} has determinant 1 and inverse $A_{-\theta}$. It is an orthogonal matrix; it preserves lengths. The joint density of (W_1, W_2) is

$$\frac{1}{2\pi} \exp\left(-\|(w_1, w_2)A^{-1}\|^2/2\right) = \frac{1}{2\pi} \exp\left(-(w_1^2 + w_2^2)/2\right)$$



A more intuitive explanation is based on the approximation

 $\mathbb{P}\{\mathbf{Z} \in B\} \approx f(\mathbf{z}) (\text{volume of } B)$

for a small ball B centered at z. The transformation from Z to W corresponds to a rotation, so

$$\mathbb{P}\{\mathbf{Z}\in B\}=\mathbb{P}\{\mathbf{W}\in B^*\},\$$

where B^* is a ball of the same radius, but centered at the point $\mathbf{w} = (w_1, \ldots, w_n)$ for which $w_1\mathbf{q}_1 + \ldots + w_n\mathbf{q}_n = \mathbf{z}$. The last equality implies $\|\mathbf{w}\| = \|\mathbf{z}\|$, from which we get

$$\mathbb{P}\{\mathbf{W}\in B^*\}\approx (2\pi)^{-n/2}\exp(-\frac{1}{2}\|\mathbf{w}\|^2) \text{ (volume of } B^*\text{)}.$$

That is, W has the asserted spherical normal density.

<12.6> **Definition.** Let $\mathbf{Z} = (Z_1, Z_2, ..., Z_n)$ have a spherical normal distribution $N(\mathbf{0}, I_n)$. The CHI-SQUARE, χ_n^2 , is defined as the distribution of $\|\mathbf{Z}\|^2 = Z_1^2 + ... + Z_n^2$.

To prove results about the spherical normal it is often merely a matter of transforming to an appropriate orthonormal basis.

<12.7> **Exercise.** Suppose
$$Z_1, Z_2, \ldots, Z_n$$
 are independent, each distributed $N(0, 1)$. Define

$$\overline{Z} = \frac{Z_1 + \ldots + Z_n}{n}$$
 and $T = \sum_{i \le n} (Z_i - \overline{Z})^2$

Show that \overline{Z} has a N(0, 1/n) distribution independently of T, which has a χ^2_{n-1} distribution.

chi-square

© David Pollard

SOLUTION: Choose the new orthonormal basis with $\mathbf{q}_1 = (1, 1, ..., 1)'/\sqrt{n}$. Choose $\mathbf{q}_2, ..., \mathbf{q}_n$ however you like, provided they are orthogonal unit vectors, all orthogonal to \mathbf{q}_1 . In the new coordinate system,

$$\mathbf{Z} = W_1 \mathbf{q}_1 + \ldots + W_n \mathbf{q}_n$$

We could calculate each W_i by dotting the sum on the right- hand side with \mathbf{q}_i : only W_i would survive. In particular,

$$W_1 = \mathbf{Z} \cdot \mathbf{q}_1 = \frac{Z_1 + \ldots + Z_n}{\sqrt{n}} = \sqrt{n}\overline{Z}.$$

From Theorem <12.5> we know that W_1 has a N(0, 1) distribution. It follows that \overline{Z} has a N(0, 1/n) distribution.

The random variable T equals the squared length of the vector

$$(Z_1-\bar{Z},\ldots,Z_n-\bar{Z})=\mathbf{Z}-\bar{Z}(\sqrt{n}\mathbf{q}_1)=\mathbf{Z}-W_1\mathbf{q}_1=W_2\mathbf{q}_2+\ldots+W_n\mathbf{q}_n$$

That is,

$$T = \|W_2 \mathbf{q}_2 + \ldots + W_n \mathbf{q}_n\|^2 = W_2^2 + \ldots + W_n^2,$$

a sum of squares of n - 1 independent N(), 1 random variables, which has a χ^2_{n-1} -distribution.

Finally, notice that \overline{Z} is a function of W_1 , whereas T is a function of the independent random variables W_2, \ldots, W_n . The independence of \overline{Z} and T follows.

Exercise for the reader: Suppose $X_1, \ldots X_n$ are independent, each distributed $N(\mu, \sigma^2)$. Apply the results from the last Exercise, with $Z_i = (X_i - \mu)/\sigma$, to deduce that \bar{X} is distributed $N(\mu, \sigma^2/n)$ independently of

$$\sum_{i\leq n} (X_i - \bar{X})^2 / \sigma^2,$$

which has a χ^2_{n-1} distribution.